BİLGİSAYAR PROJESİ - II
YASİN SEZGİN
****

ONLINE SATIŞ SİTELERİNDE YAPILAN YORUMLARIN 1 PUAN İLE 5 PUAN ARASINDA SINIFLANDIRILMASI

Problem

Online alışveriş sitelerinde ürünlere yapılan yorumlar olası diğer alıcılara yardımcı olur mu?

Online alışveriş sitelerinde ürünlere yapılan yorumlar ürün ve hizmet kalitesinin artırılmasında rol oynayabilir mi?

Online alışveriş sitelerinde ürünlere yapılan yorumlar doğru mu?

Projenin Amacı

Online alışveriş sitelerinde yapılan yorumların ayrıştırılması ve yapılan değerlendirmeler ile alaka düzeyinin makine öğrenmesi yöntemleri (SVM, Naive Bayes vb.) ile sınıflandırılması ve ilgili yorumların ilgili puanlara otomatik bir biçimde atanması amaçlanmaktadır.

Metin Sınıflandırma

Günümüzde Google, Amazon gibi dünyanın en büyük şirketleri makine öğrenmesi ve derin öğrenme alanlarına büyük yatırımlar yapmaktadır.

Makine öğrenmesi ve Veri Madenciliği ve Büyük Veri çalışmalarında sıkça kullanılmaktadır.

Metin sınıflandırması metinlerin içeriklerine göre belirli sınıflar altında hangi sınıfa ait ise o sınıfa atanması işlemidir.

Bir metni belirlenmiş kategorilere atama, spam filtreleme, duygu analizi yapma birer metin sınıflandırma görevi olabilirler.

Çalışma Ortamı ve Kullanılan Teknolojiler

Python Programlama Dili

Jupyter Geliştirme Ortamı

Numpy, Pandas, Scikit-Learn, Matplotlib vb. python kütüphaneleri

Makine Öğrenmesi Uygulama

1. Verisetinin elde edilmesi
2. Veriseti Ön işleme
3. Modelleri üretme ve eğitim
4. Modelleri test etme
5. Modellerin değerlendirilmesi

Veriseti

Bu veri seti, popüler Türk alışveriş sitesi hepsiburada.com'dan toplanmıştır. Yıldız derecelendirmeleriyle birlikte toplam 272.216 yorum içerir.

https://www.kaggle.com/cebeci/turkishreviews

	Rating	Review	URL
0	5	3 yıldır tık demedi. :)	https://www.hepsiburada.com/logitech-m175-kabl...
1	5	3 yıldır kullanıyorum müthiş	https://www.hepsiburada.com/logitech-m175-kabl...
2	4	Ürün bugün elime geçti çok fazla inceleme fırs...	https://www.hepsiburada.com/logitech-m175-kabl...
3	4	Almaya karar verdim. Hemencecik geldi. Keyifle...	https://www.hepsiburada.com/logitech-m175-kabl...
4	5	Günlük kullanımınızı çok çok iyi karsılıyor kı...	https://www.hepsiburada.com/logitech-m175-kabl...
...	...	...	...
272211	5	fiyatına göre güzel	https://www.hepsiburada.com/samsung-galaxy-gra...
272212	5	Ürün kullanışlı iş görüyor fazlasıyla eşime al...	https://www.hepsiburada.com/samsung-galaxy-gra...
272213	5	Hızlı Kargo, güzel ürün	https://www.hepsiburada.com/samsung-galaxy-gra...
272214	5	telefon başarılı hızlı bir cihaz sadece beyaz...	https://www.hepsiburada.com/samsung-galaxy-gra...
272215	4	Urun cok guzel pazar gunu siparis verdim adana...	https://www.hepsiburada.com/samsung-galaxy-gra...

272216 rows × 3 columns

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 272216 entries, 0 to 272215
Data columns (total 3 columns):
 #   Column  Non-Null Count   Dtype 
---  ------  --------------   ----- 
 0   Rating  272216 non-null  int64 
 1   Review  272216 non-null  object
 2   URL     272216 non-null  object
dtypes: int64(1), object(2)
memory usage: 6.2+ MB
```

Veriseti

‘‘3 yıldır tık demedi. :)’’

‘‘3 yıldır kullanıyorum müthiş’’

‘‘Ürün bugün elime geçti çok fazla inceleme fırs...’’

…

Verisetinin ilk hali görüldüğü gibi temizlenmemiş, doğal bir şekildedir ve bu haliyle herhangi bir öğrenme algoritması için hazır değildir.

Veriseti Ön İşleme

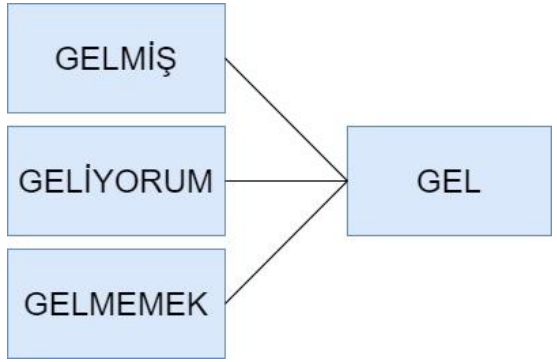
Metinleri makineye anlatabilmek için belirli şablonlar haline getirmeliyiz.

1. Metnin küçük harflere indirgenmesi
2. Metinden mentionların, URL'lerin, noktalama işaretlerinin ve emojilerin çıkartılması
3. Metindeki kelimelerin kelime köklerine indirgenmesi (Lemmatisation)
4. Metinden stopwordlerin (anlama etki etmeyen kelimeler: ve, ile vb.) çıkartılması

Veriseti Ön İşleme Lemmatisation

Lemmatisation, yani kelimenin köklerine indirgenmesi işlemidir. (Özellikle Türkçe gibi sondan eklemeli dillerde bu indirgeme işlemi çok önemlidir.)

Kelimelerin kelime köklerine indirgenmesi için Zemberek-NLP kullanıldı. Zemberek-NLP, Türkçe için Doğal Dil İşleme sağlar.



https://github.com/ahmetaa/zemberek-nlp

Veriseti

Gerekli ön işleme adımları izlendikten sonra verisetinin son hali bu şekildedir.

	Rating	Review	URL
0	5	yıl tık de	https://www.hepsiburada.com/logitech-m175-kabl...
1	5	yıl kullan müthiş	https://www.hepsiburada.com/logitech-m175-kabl...
2	4	ürün bugün el geç çok fazla incele fırsat ol g...	https://www.hepsiburada.com/logitech-m175-kabl...
3	4	al karar ver hemencecik gel keyif kullan	https://www.hepsiburada.com/logitech-m175-kabl...
4	5	günlük kullanım çok çok iyi karsılıyor kısaç m...	https://www.hepsiburada.com/logitech-m175-kabl...
...	...	...	...
272211	5	fiyat göre güzel	https://www.hepsiburada.com/samsung-galaxy-gra...
272212	5	ürün kullan iş gör fazlasıyla eş aldi cok memn...	https://www.hepsiburada.com/samsung-galaxy-gra...
272213	5	hız kargo güzel ürün	https://www.hepsiburada.com/samsung-galaxy-gra...
272214	5	telefon başarı hız cihaz beyaz iste gri renk gel	https://www.hepsiburada.com/samsung-galaxy-gra...
272215	4	ur cok guzel pazar gun siparis ver adana sali ...	https://www.hepsiburada.com/samsung-galaxy-gra...

272216 rows × 3 columns

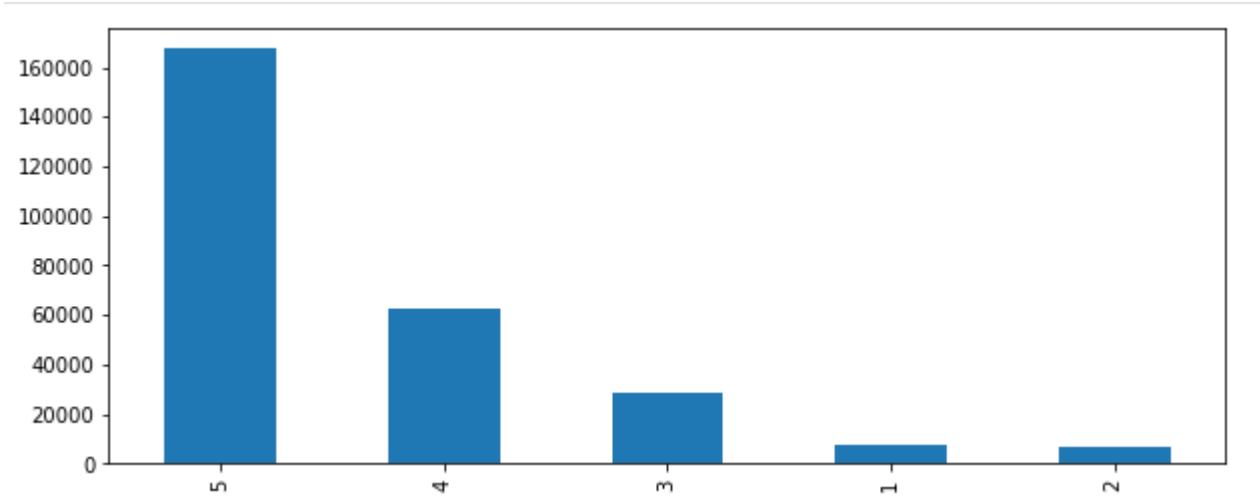
Veriseti Ön İşleme Vektörizasyon

Metin Vektörleştirme, metni sayısal gösterime dönüştürme işlemidir. Metinleri makineye anlatabilmek için belirli kalıplarda vektörizasyon yapılmaktadır.

Bu işlem için TfidfVectorizer kullanıldı.

Veriseti

Veriseti sınıf dağılımı olarak dengesizdir. 272.216 yorumun yaklaşık 160.000'i 5 yıldız yorumlarından oluşmaktadır.



Modellerin değerlendirilmesi

Accuracy: Doğru etiketle kategorilere ayrılmış metinlerin yüzdesi.

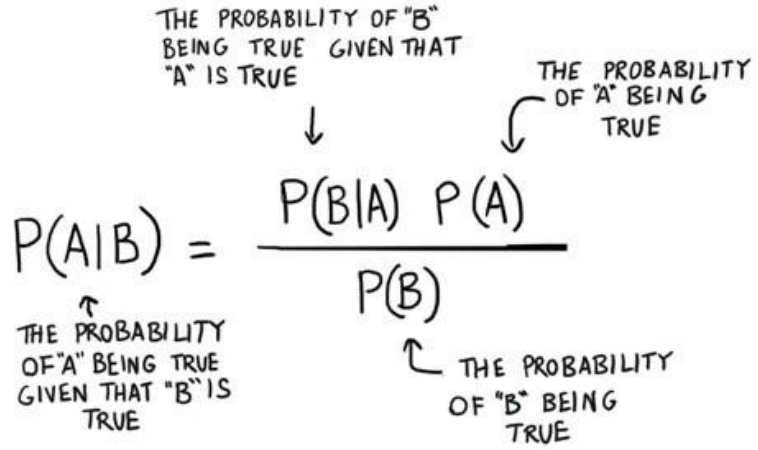
Precision: Sınıflandırıcının belirli bir etiket için tahmin ettiği toplam örnek sayısından elde ettiği örneklerin yüzdesi.

Recall: Sınıflandırıcının, belirli bir etiket için tahmin etmesi gereken toplam örnek sayısı içinden belirli bir etiket için tahmin ettiği örneklerin yüzdesi.

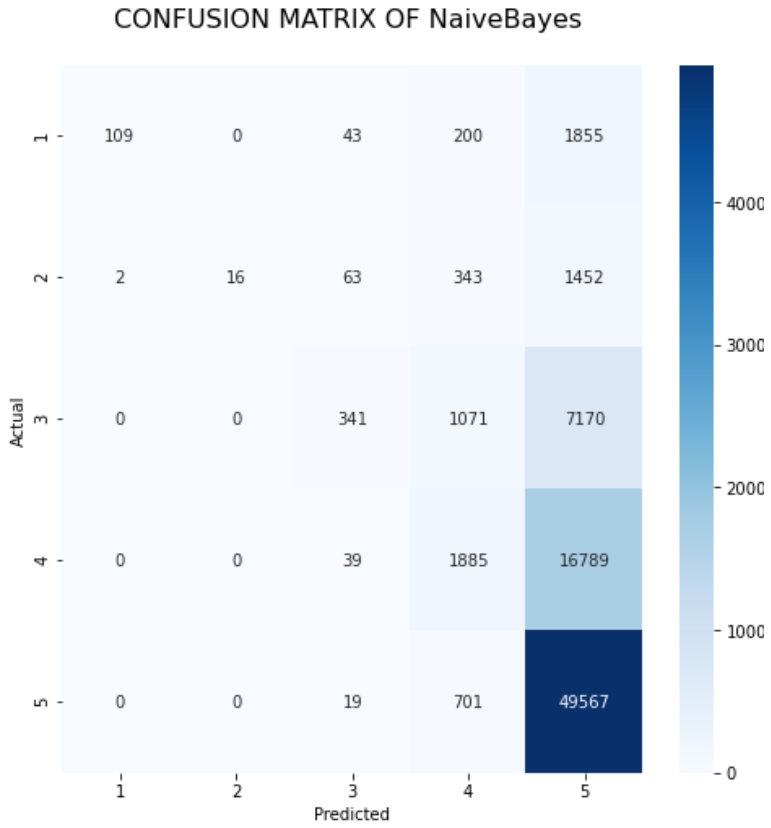
F1 Score: Kesinlik ve geri çağırmanın harmonik ortalaması.

Modeli üretme ve eğitim
Naive Bayes

Naïve Bayes sınıflandırıcı, örüntü tanıma problemine ilk bakışta oldukça kısıtlayıcı görülen bir önerme ile kullanılabilen olasılıksal bir yaklaşımdır. Bu önerme, örüntü tanımada kullanılacak her bir tanımlayıcı öznitelik ya da parametrenin istatistik açıdan bağımsız olması gerekliliğidir. Her ne kadar bu önerme sınıflandırıcının kullanım alanını kısıtlasa da istatistik bağımsızlık koşulu esnetilerek kullanıldığında da daha karmaşık yapay sinir ağları gibi metotlarla karşılaştırabilir sonuçlar vermektedir. Bir Naive Bayes sınıflandırıcı, her özniteliğin birbirinden koşulsal bağımsız olduğu ve öğrenilmek istenen kavramın tüm bu özniteliklere koşulsal bağlı olduğu bir Bayes ağı olarak da düşünülebilir.



Modeli test etme
Naive Bayes

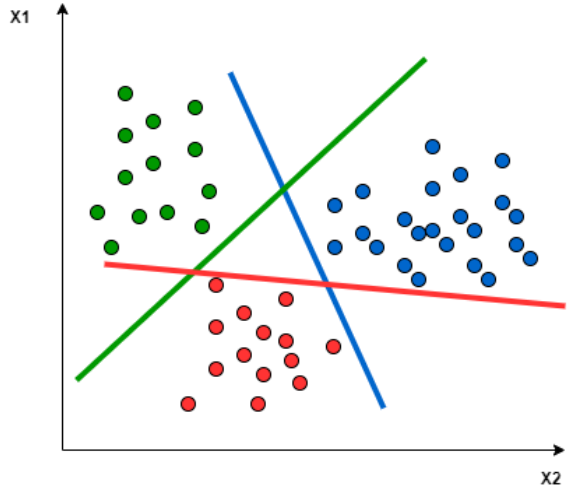


Modeli test etme
Naive Bayes

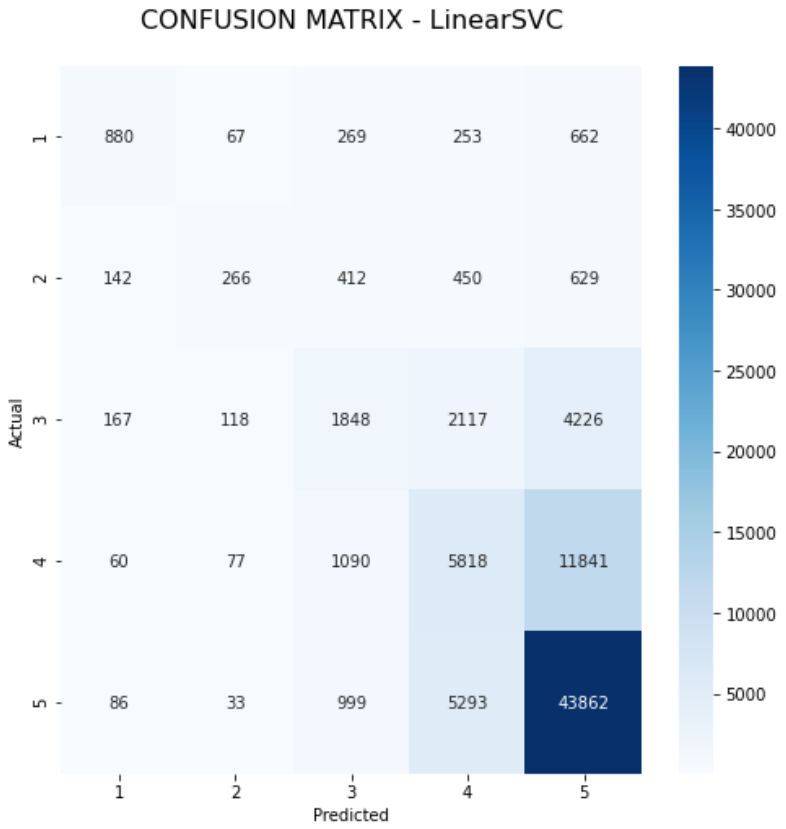
	precision	recall	f1-score	support
1	0.98	0.05	0.09	2207
2	1.00	0.01	0.02	1876
3	0.68	0.04	0.08	8582
4	0.45	0.10	0.16	18713
5	0.65	0.99	0.78	50287
accuracy			0.64	81665
macro avg	0.75	0.24	0.23	81665
weighted avg	0.62	0.64	0.53	81665

Model üretme ve eğitim
Linear SVM

Destek Vektör Makineleri, sınıflandırma problemleri için kullanılabilen denetimli bir makine öğrenmesi algoritmasıdır. Bu algoritmada, her bir veri maddesini belirli bir koordinatın değeri olan her özelliğin değeri ile birlikte n-boyutlu boşluğa (burada n sahip olduğunuz özelliklerin sayısı) bir nokta olarak çizilir. Ardından, iki sınıftan oldukça iyi ayrım yapan hiper-düzlemi bularak sınıflandırma gerçekleştirilir.



Modeli test etme
Linear SVM

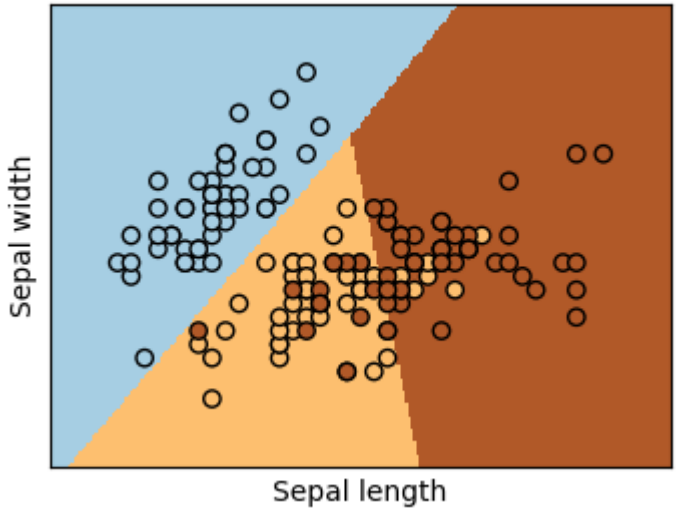


Modeli test etme
Linear SVM

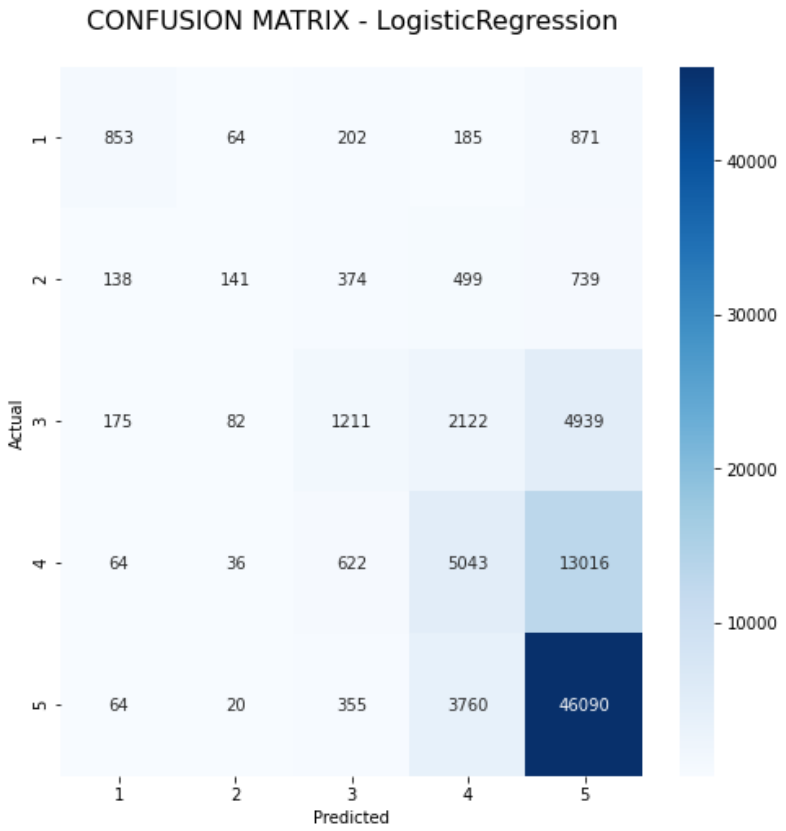
	precision	recall	f1-score	support
1	0.66	0.41	0.51	2131
2	0.47	0.14	0.22	1899
3	0.40	0.22	0.28	8476
4	0.42	0.31	0.35	18886
5	0.72	0.87	0.79	50273
accuracy			0.65	81665
macro avg	0.53	0.39	0.43	81665
weighted avg	0.61	0.65	0.61	81665

Model üretme ve eğitim
Logistic Regression

Lojistik regresyon, bir sonucu belirleyen bir veya daha fazla bağımsız değişken bulunan bir veri kümesini analiz etmek için kullanılan istatistiksel bir yöntemdir.



Modeli test etme
Logistic Regression



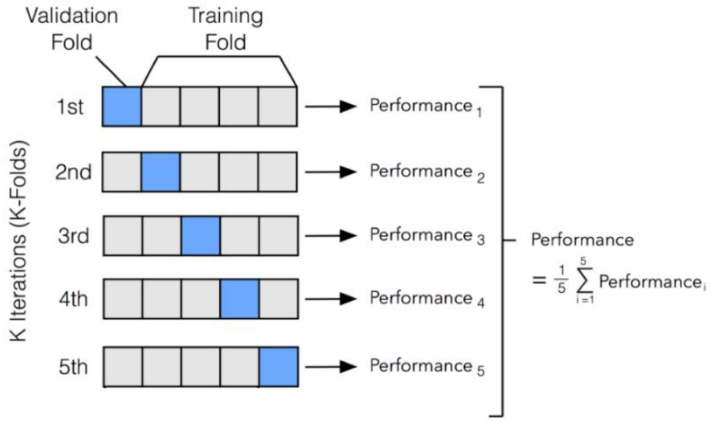
Modeli test etme
Logistic Regression

CLASSIFICATIION METRICS

	precision	recall	f1-score	support
1	0.66	0.39	0.49	2175
2	0.41	0.07	0.13	1891
3	0.44	0.14	0.21	8529
4	0.43	0.27	0.33	18781
5	0.70	0.92	0.80	50289
accuracy			0.65	81665
macro avg	0.53	0.36	0.39	81665
weighted avg	0.61	0.65	0.60	81665

Modelleri karşılaştırma Cross-Validation

Cross-validation, makine öğrenmesi modelinin görmediği veriler üzerindeki performansını mümkün olduğunca objektif ve doğru bir şekilde değerlendirmek için kullanılan istatistiksel bir yeniden örnekleme yöntemidir.



Modelleri karşılaştırma

```
accuracy of naive bayes         with 5 fold cross validation  0.5958906621192384
accuracy of linear svm          with 5 fold cross validation  0.577838793552281
accuracy of logistic regression with 5 fold cross validation  0.6021209960049413
```

Dinlediğiniz için teşekkürler.